CaTER: A Framework for Context-aware Topology Entity Retrieval Contrastive Learning in End-to-End Task-Oriented Dialogue Systems

Di Wu and Zhizhi Yu *
Hebei University of Engineering School of Information and Electronic Engineering

Abstract

Retrieving entity knowledge that aligns with user intent is essential for task-oriented dialogue (TOD) systems to support personalization and localization, especially under large-scale knowledge bases. However, generative models tend to suffer from implicit association preference, while retrieval-generation approaches face knowledge transfer discrepancies. To address these challenges, we propose CaTER, a Context-aware Topology Entity Retrieval Contrastive Learning Framework. CaTER introduces a cycle context-aware distilling attention mechanism, which employs context-independent sparse pooling to suppress noise from weakly relevant attributes. We further construct topologically hard negative samples by decoupling entity information from generated responses and design a topology entity retrieval contrastive loss to train the retriever by reverse distillation. Extensive experiments on three standard TOD benchmarks with both small and large-scale knowledge bases show that CaTER consistently outperforms strong baselines such as MAKER and MK-TOD, achieving state-of-the-art performance in TOD system.

1 Introduction

Task-Oriented Dialogue (TOD) systems must effectively handle localized and personalized tasks by retrieving knowledge about specific entities from large-scale Knowledge Base (KB). The rapid advancement of large language models (Yang et al., 2024) such as deepseek (Liu et al., 2024) and ChatGPT (OpenAI, 2022) has introduced novel paradigms for TOD systems in large-scale KB scenario, thereby significantly accelerating the development of TOD technologies.

One key limitation of current TOD systems is the lack of belief state supervision, which traditionally guides accurate retrieval from external KBs. Existing methods can be broadly categorized into two paradigms. The first line of work(Huang et al., 2022; Madotto et al., 2020) integrates entity retrieval into response generation, enabling implicit knowledge retrieval by encoding KB facts into model parameters under reference-response supervision. The second line (Rony et al., 2022; Xie et al., 2022; Tian et al., 2022) decouples retrieval and generation, using pseudo-supervision from generated responses to improve entity selection. For example, Q-TOD (Tian et al., 2022) extracts key information from dialogue context and explicitly retrieves relevant entities before generating responses.

Figure 1: An example of entity knowledge retrieval. The retriever equally considers both highly relevant attributes (marked in red) and weakly relevant attributes (marked in green), inducing retrieval-generation bias and resulting in incorrect dialogue responses.

While some approaches adopt semi-supervised training strategies (e.g., GALAXY (He et al., 2022), LABES (Zhang et al., 2020), JSA-KRTOD (Cai et al., 2023)), we consider supervision level to be orthogonal to retrieval architecture: both paradigms can be trained in a fully- or semi-supervised way.

The Distractive Attribution Problem (DAP) first formalized by ReAL (Chen et al., 2024), which highlighted that false but similar knowledge, such as hard negative entities, can mislead generation models due to attribute distraction. ReAL proposes

*Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11633–11648
November 4-9, 2025

a two-stage contrastive framework to alleviate DAP through adaptive negative sampling and generator-guided retriever alignment.

Building upon this insight, we further analyze DAP from a fine-grained modeling perspective, identifying three key error sources that jointly degrade TOD performance, as illustrated in Figure 1: (1)**Entity Retrieval Inaccuracy**. TOD systems are prone to errors when retrieving Top-K entities from the KB, as the process is often affected by ambiguous entity attributes in the dialogue context. (2)**Implicit Association Preference**. The attribute similarity among Top-K retrieved entities tends to induce retrieval-generation bias, causing the generator to over-rely on high-frequency entities in the training corpus and produce incorrect dialogue responses. (3)**Retrieval-Generation Knowledge Transfer Discrepancy**. Excessive attribute similarity among Top-K retrieved entities limits the generator's ability to make fine-grained distinctions, leading to incorrect responses. These errors introduce noisy supervision signals, which in turn back-propagate flawed knowledge to the retriever and amplify bias through a reinforcement loop.

Specifically, we define Retrieval-Generation Knowledge Transfer Discrepancy as the misalignment between retrieved knowledge and the factual content grounded in the generated response. While the retriever may select entities whose attributes are superficially relevant, the generator may misattribute or blend information from incorrect entities due to high attribute similarity and insufficient grounding supervision.

This discrepancy is distinct from retrieval inaccuracy (incorrect entities retrieved) and implicit association bias (generator favoring popular entities), as it arises after seemingly reasonable retrieval, but results in hallucinated or mixed-slot responses. A concrete example is discussed in 1, where CaTER retrieves the correct restaurant ("hakka") but generates an incorrect area information from a hard negative ("hk fusion"). The same phenomenon was happened in the failure case at Appendix C.

To address these issues, we propose a **C**ontext-**a**ware **T**opology **E**ntity **R**etrieval Contrastive Learning framework (**CaTER**), which introduces a Cycle Context-Aware Distilling Attention (CyCAD) mechanism that performs context-independent sparse entity pooling, allowing the generator to focus more on entity knowledge and reduce interference from weakly relevant attributes. Additionally, we design a Topology Entity Retrieval Contrastive Learning method (TER) that constructs topologically hard negative samples during training and optimizes a topological contrastive loss to reduce generator bias toward head entities.

Extensive experiments on three standard TOD benchmarks, MultiWOZ 2.1, CamRest, and SMD, demonstrate that CaTER consistently outperforms strong baselines in both dialogue-level and dataset-level KB settings.

Our main contributions are as follows:

- We propose CaTER, a novel context-aware contrastive learning framework that jointly improves entity retrieval and response generation.
- We introduce TER to construct topological hard negatives and optimize the retriever by contrastive loss under reverse distillation.
- We validate CaTER on three TOD benchmarks, achieving state-of-the-art performance across both small-scale and large-scale KB settings.

2 Related Work

2.1 End-to-End Task-Oriented Dialogue

Early studies on end-to-end TOD systems performed implicit retrieval by leveraging KB within generator parameters during response generation. KE (Madotto et al., 2020) proposed a method that directly incorporates the KB into the parameters of the backbone model, that generates equivalent dialogue responses based on the user query and entity information, while implicitly storing KB content. ECO (Huang et al., 2022) introduced an end-to-end TOD system that employs trie-constrained autoregressive generation to produce the most relevant entities, ensuring consistency in entity usage within generated responses. Similarly, Simple-TOD (Ding et al., 2024) performs implicit knowledge retrieval by generating entities through multi-level prefix knowledge tries with autoregressive decoding. The above methods simplify the architecture of TOD systems and effectively address issues such as error propagation in pipeline paradigms and data dependency. However, as the KB grows, end-to-end retrieval-generation models face increased difficulty in selecting accurate knowledge due to the entanglement of retrieval objectives with generation loss, as observed in prior studies (Lewis et al., 2020).

Compared to traditional explicit retrieval methods, recent researches has increasingly focused on optimizing the architecture of end-to-end TOD systems to enhance their explicit entity retrieval capabilities. Q-TOD (Tian et al., 2022) decouples knowledge retrieval from response generation by extracting salient information from the dialogue context to retrieve relevant knowledge for response generation. MAKER (Wan et al., 2023) filters user-intent-aligned entities through both entity selector and attribute selector. DF-TOD (Shi et al., 2023) extracts supervision signals from generated responses and uses them as pseudo-labels to train the retriever. MK-TOD (Shen et al., 2023) leverages meta-knowledge to guide generator training. Supervision signals are extracted from dialogue responses, and the retriever is trained in reverse by marginal likelihood maximization. An approach (Xu et al., 2024) for knowledge retrieval driven by matching representations is proposed, that performs entity re-ranking through matching signal extraction and attribute-based filtering. The above studies demonstrate that building deeply integrated frameworks can effectively address the challenge of fine-grained entity knowledge matching. However, they still face limitations in mitigating knowledge transfer discrepancy between retrieval-generation.

2.2 Contrastive Learning

Contrastive learning has demonstrated strong representational power in dialogue systems, enhancing intent understanding, cross-modal alignment, and response generation by modeling semantically discriminative spaces. It also provides a new perspective for unifying entity retrieval and response generation.

In retrieval-based systems, DPR (Karpukhin et al., 2020) introduced a dual-encoder architecture trained with contrastive loss to distinguish relevant from irrelevant passages, laying the foundation for contrastive dense retrieval in both QA and dialogue settings. Follow-up works have extended this idea to task-oriented dialogue, using contrastive objectives to improve entity retrieval or response quality.

For example, Dial2vec (Liu et al., 2022) captures interaction patterns between dialogue participants to learn distinct embeddings. In the domain of visual dialogue. Utc (Chen et al., 2022a) leverages a context contrastive loss and an answer contrastive loss to provide representation learning signals from different perspectives. ICMU (Chen et al., 2022b) enhances cross-modal comprehension by distinguishing different retrieved inputs through a four-way contrastive learning strategy. In the field of dialogue generation, a contrastive learning method is proposed (Tan et al., 2023), constructs contrastive samples based on salient dialogue attributes to enhance response generation. An entity-based contrastive learning framework (Wang et al., 2024) leverages entity information from dialogue samples to construct positive and negative examples, involving semantically relevant and irrelevant perturbations, respectively. The above studies have significantly improved the semantic representation capabilities and controllability of TOD systems through the design of contrastive learning methods.

Inspired by the above researches, this paper introduces contrastive learning into TOD systems and designs a topology entity retrieval contrastive learning framework.

3 The CaTER Framework

A novel CaTER framework is proposed, as illustrated in Figure 2. Relevant entity attributes are masked through entity and attribute score. A CyCAD is introduced, which employs context-independent sparse entity pooling to mask dialogue context information, guiding the generator to focus more on entity knowledge. A TER method is designed, which leverages CyCAD during training to construct topologically hard negative entity samples. Response entity is decoupled from dialogue responses, and a topology entity contrastive loss is introduced. The retriever is then trained by reverse distillation.

3.1 Preliminaries

Given a dialogue $\mathcal{D} = \{u_1, y_1, ..., y_{T-1}, u_T\}$, where u_t and y_t are the t-th turn user utterance and system response, respectively, T denotes the T-th dialogue turn. c_t denotes the dialogue context at the t-th turn, defined as a subset of $\mathcal{D} = u_1, y_1, \ldots, u_{t-1}, y_{t-1}, u_t$. In addition, to accommodate domain-specific requirements, an external KB $\mathcal{K} = \{e_1, ..., e_i, ..., e_B\}$ is provided in the form of a set of B entities. Each entity e_i is composed of N attribute–value pairs It denoted as $e_i = \{a^1, v_i^1, ..., a^N, v_i^N\}$. An end-to-end TOD system takes c_t and $\mathcal{K}$ as input, and directly generates a natural language response $\mathcal{R}$.

3.2 Multi-Grained Entity Retrieval

Inspired by (Wan et al., 2023), the proposed multi-grained entity retrieval stage incorporates both en-

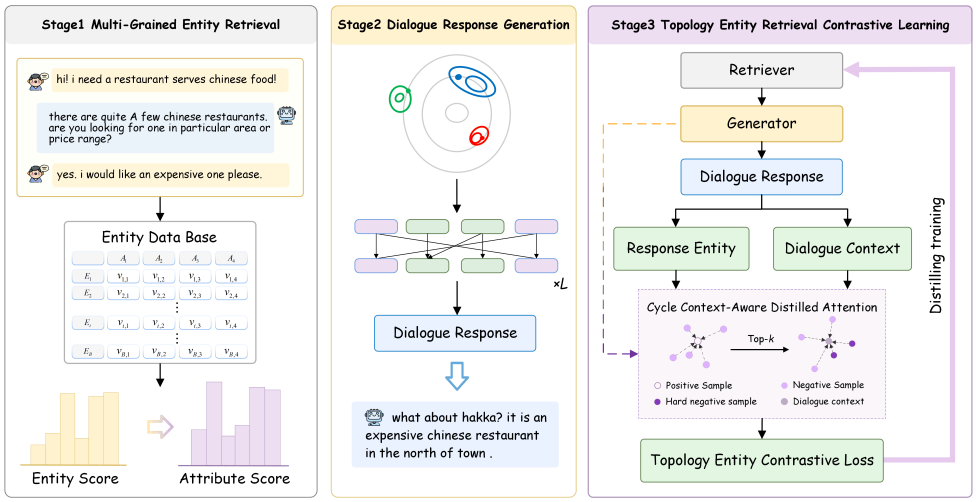

Figure 2: Architecture of the CaTER.

tity and attribute score to select entities aligned with user intent. To ensure simplicity and effectiveness, a dual-encoder architecture is adopted. One encoder Enc_c is used to concatenate and encode u_t and y_t , as the query, while the other encoder Enc_e encodes the entity information from $\mathcal{K}$.

Specifically, entity information is encoded by concatenating its attribute–value pairs and feeding the resulting sequence into Enc_e. The entity score $s_{t,i}$ for e_i is computed as the dot product between c_t and the e_i as following:

$$s_{t,i} = (\mathrm{Enc}_c(c_t))^T \mathrm{Enc}_e(e_i)$$

where $\mathrm{Enc}_c(\cdot)$ represents the context encoder Enc_c, $\mathrm{Enc}_e(\cdot)$ denotes the entity encoder Enc_e, $(\cdot)^T$ is the transpose operation.

Based on $s_{t,i}$, the Top-K entities are initially selected from $\mathcal{K}$ as candidate entities for response generation, denoted as $\hat{\mathcal{K}} = \{e_1, e_2, ..., e_K\}$.

Pre-trained language model BERT (Devlin et al., 2019) is used to construct the encoders, where the representation of the [CLS] token is extracted to represent c_t and e_i. Existing studies (Qin et al., 2019) have emphasized that directly initializing encoders with BERT weights may lead to representation collapse, which degrades retrieval performance.

Therefore, following the approach proposed in (Shi et al., 2023). In this paper, we adopts distant supervision-based pretraining to initialize the retriever weights.

In addition, an attribute score is designed to remove irrelevant attribute–value pairs from the retrieved entities. c_t is concatenated with each entity $e_i \in \hat{\mathcal{K}}$, and the resulting sequence is encoded by an attribute encoder Enc_a built on a pre-trained language model. The final [CLS] token from Enc_a is extracted and passed through a Feed-Forward Network (FFN) to obtain an N-dimensional vector, which is used to compute the attribute score $\mathbf{a}_{t,i}$ as following:

$$\mathbf{a}_{t,i} = \mathrm{FFN}(\mathrm{Enc}_a([c_t; e_i]))$$

where $\mathbf{a}_{t,i} \in \mathbb{R}^N$ represents the importance of the corresponding attribute.

The $\mathbf{a}_{t,i}$ of all retrieved entities are summed, and a cumulative importance score $\mathbf{a}_t$ is computed by weighting them with the corresponding $s_{t,i}$ as following:

$$\mathbf{a}_t = \sigma(\sum_{i=1}^{K} s_{t,i} \cdot \mathbf{a}_{t,i})$$

where $\sigma(\cdot)$ denotes the Sigmoid function.

Attributes with scores greater than a predefined threshold τ_a are selected, and an attribute subset $\tilde{\mathcal{K}} = \{\tilde{e}_1, ..., \tilde{e}_K\}$ is constructed by masking irrelevant attribute–value pairs from each retrieved entity in the candidate entity set $\hat{\mathcal{K}}$.

A discrete N-dimensional 0-1 vector $\mathbf{b}_t$ is constructed based on whether each masked attribute value in the $\hat{\mathcal{K}}$ appears in c_t. It is used to define a context-aware attribute score loss $\mathcal{L}_{attr}$ as following:

$$\mathcal{L}_{attr} = \mathrm{BCELoss}(\mathbf{a}_t, \mathbf{b}_t)$$

In addition, pre-trained language models are used to implement the encoders Enc_c, Enc_e and Enc_a. Following the approach in (Wan et al., 2023), the pre-trained models are initialized accordingly, and the final [CLS] token representation is used as the encoder output.

3.3 Dialogue Response Generation

In this section, a Cycle Context-Aware Distillation Attention (CyCAD) is designed to explicitly reduce the interference of weakly correlated attributes on the generator through context-independent sparse entity pooling and context-aware attention fusion, which effectively improves the accuracy of dialogue responses and the consistency of entity knowledge.

A novel attention mechanism CyCAD is constructed to mitigate the influence of weakly relevant attributes through context-independent sparse entity pooling.

The generator is built upon the pre-trained T5 model to facilitate direct interaction between c_t and $\hat{\mathcal{K}}$. It primarily consists of an encoder Enc_g and a decoder Dec_g. Specifically, c_t and the $\tilde{e}_k$ are concatenated and independently processed by Enc_g to construct a global representation for the current dialogue turn $\mathbf{H}_{t,i}$ as following:

$$\mathbf{H}_{t,i} = \mathrm{Enc}_g([c_t, \tilde{e}_k])$$

The representations of all retrieved entities are concatenated and used as the input $\mathbf{H}_t$ to Dec_g.

Dec_g generates dialogue responses in an autoregressive manner. Specifically, context-independent entity pooling is constructed to filter out irrelevant contextual information, enabling the generator to focus on distilling entity knowledge. Context-aware average pooling is used to compute $Pool^c_{t,i}$ as following:

$$Pool^c_{t,i} = \frac{\sum_{m=1}^{L_c} Att_c[t,i,m] \cdot \mathbf{M}_{mask}[t,i,m]}{\sum_{m=1}^{L_c} \mathbf{M}_{mask}[t,i,m] + \epsilon}$$

where L_c represents the length of c_t, $Att_c[t,i,m]$ and $\mathbf{M}_{mask}[t,i,m]$ denote the cross-attention score of the *m*-th token at *t*-th dialogue turn with respect to the masked candidate entity $\tilde{e}_i$, and the corresponding input attention mask matrix. ϵ is the bias.

To eliminate the influence of weakly relevant attributes in c_t on entity retrieval, context-independent entity pooling is performed solely based on entity representations $Pool^e_{t,i}$ as following:

$$Pool^e_{t,i} = \frac{\sum_{m=L_c+1}^{L_{\mathbf{H}_t}} Att_c[t,i,m] \cdot \mathbf{M}_{mask}[t,i,m]}{\sum_{m=L_c+1}^{L_{\mathbf{H}_t}} \mathbf{M}_{mask}[t,i,m] + \epsilon}$$

The CyCAD score $s_{att}(c_t, e_i)$ is computed by controlling the information fusion ratio as following:

$$s_{att}(c_t, e_i) = \alpha \cdot Pool^c_{t,i} + (1-\alpha) \cdot Pool^e_{t,i}$$

where α represents the context-aware pooling weight parameter.

The $s_{att}(c_t, e_i)$ is normalized using a softmax function to obtain the CyCAD distribution $\hat{\mathbf{S}}_{att}$ over the Top-*K* retrieved entities, reflecting their importance in the generating response.

The system response token probability distribution $P(\mathcal{R}_{t,r})$ is generated based on $\mathbf{H}_t$ as following:

$$P(\mathcal{R}_{t,r}) = \mathrm{Dec}_g(\mathcal{R}_{t,r} | \mathcal{R}_{t,<r}, \mathbf{H}_t)$$

where $\mathcal{R}_{t,i}$ represents the *i*-th token in the generated response at *t*-th dialogue turn, r denotes the max length of generate dialogue response.

The generator is trained using the standard cross-entropy loss $\mathcal{L}_{gen}$ as following:

$$\mathcal{L}_{gen} = \sum_{r=1}^{r} -\log P(\mathcal{R}_{t,r})$$

3.4 Topology Entity Retrieval Contrastive Learning

In this section, we propose the Topology-aware Entity Retrieval Contrastive Learning method (TER), which leverages the CyCAD to generate topological entity hard-negative samples, explicitly distinguishes fine-grained entity differences through contrastive learning loss, and achieves the collaborative training of the retriever and the generator through reverse distillation, which significantly enhances the backbone model's ability to distinguish fine-grained entities. The illustration of TER is showed in Figure 3.

The response is decoupled to isolate entity information, and CyCAD is employed to compute the entity relevance score $S_{att}(\mathcal{R}, e_i)$ between entities and the dialogue context as following:

$$S_{att}(\mathcal{R}, e_i) = \sum_{u=1}^{K} s_{att}(c_t, e_i)$$

Attribute values in the masked candidate entity $\hat{\mathcal{K}}$ are matched against the dialogue response, and combined with the entity relevance score $S_{att}(\mathcal{R}, e_i)$ to compute a joint entity-attribute score $\mathrm{A}_{e_i}(\mathcal{R}, e_i)$ as following:

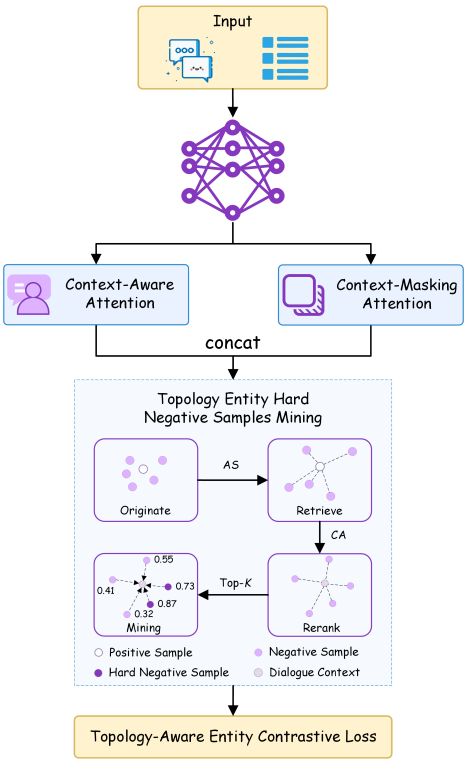

Figure 3: The illustration of topology entity retrieval contrastive learning. Where, AS denotes the attribute score $\mathbf{a}_{t,i}$, and CA denotes the cross-attention score.

$$\mathrm{A}_{e_i}(\mathcal{R}, e_i) = \sigma \sum_{i=1}^{K} S_{att}(\mathcal{R}, e_i) \cdot \mathrm{FFN}(\mathrm{Enc}_a([\mathcal{R}; e_i]))$$

The $\hat{\mathcal{K}}$ are re-ranked to construct a set of topologically hard negative samples. The calculation of positive sample d^+ selection strategy as following:

$$d^+ = \arg\max_{d \in \hat{\mathcal{K}}} \mathrm{A}_{e_i}(\mathcal{R}, d)$$

We propose an Topology Entity Hard Negative mining strategy that selects contextually plausible but ultimately incorrect entities as distractors. Specifically, for each dialog, we rank candidate entities e_i based on their attribute alignment score $\mathrm{A}_{e_i}(\mathcal{R}, e_i)$, and exclude true positives from the same domain to form hard negative samples $\mathcal{N}_{hard}$.

The calculation of topology hard negative $\mathcal{N}_{hard}$ sampling strategy as following:

$$\mathcal{N}_{hard} = \mathrm{Top} - \mathrm{Q}_{d \neq d^+}(\mathrm{A}_{e_i}(\mathcal{R}, d))$$

where $\mathrm{Top} - \mathrm{Q}_{d \neq d^+}(\cdot)$ represents the Top-Q samples d differs from the positive sample d^+.

This formulation targets cross-domain ambiguity, encouraging the model to learn more robust contrastive distinctions across structurally similar but semantically divergent entities.

In the representation learning framework of TOD systems, adversarial sample sets constructed from $\mathcal{N}_{hard}$ exhibit high-density distributions in the semantic space, with significantly lower inter-class separability compared to conventional negative samples. Based on this observation, a topology entity contrastive loss $\mathcal{L}_{cl}$ is designed by establishing a positive correlation between sample hardness and loss weight within the contrastive objective as following:

$$\mathcal{L}_{cl} = -\log \frac{e^{\frac{\mathrm{A}_{e_i}(c_t, d^+)}{\tau_t}}}{e^{\frac{\mathrm{A}_{e_i}(c_t, d^+)}{\tau_t}} + \sum_{i=1}^{K} e^{\frac{\mathrm{A}_{e_i}(c_t, \mathcal{N}_{hard})}{\tau_t}}}$$

where τ_t represents the temperature parameter.

The $\mathcal{L}_{cl}$ is normalized using a softmax function to reduce the impact of incorrectly selected entities on the overall training loss, thereby preventing the generator from learning from erroneous samples.

In addition, the Kullback–Leibler divergence between the similarity score $\mathbf{s}_t = \sum s_{t,i}$ of the retrieved entities and the CyCAD distribution is computed as the entity selection loss $\mathcal{L}_{ent}$ as following:

$$\mathcal{L}_{ent} = \mathcal{D}_{KL}(\mathbf{s}_t || \hat{\mathbf{S}}_{att})$$

Finally, the overall loss of the dialogue system $\mathcal{L}_{total}$ is defined as the sum of the attribute score loss $\mathcal{L}_{attr}$, the entity selection loss $\mathcal{L}_{ent}$, the generator training loss $\mathcal{L}_{gen}$ and the topology entity contrastive loss $\mathcal{L}_{cl}$ as following:

$$\mathcal{L}_{total} = \mathcal{L}_{attr} + \mathcal{L}_{ent} + \mathcal{L}_{gen} + \mathcal{L}_{cl}$$

4 Experiments

4.1 Datasets

We conduct experiments on three benchmark datasets: MultiWOZ 2.1 (MWOZ) (Eric et al., 2020), Stanford Multi-Domain (SMD) (Eric et al., 2017), and CamRest (Wen et al., 2017). To the best of our knowledge, each dialogue turn in these datasets is associated with a KB that contains all entity information required to fulfill the user's intent. The dataset splits follow the settings used in prior work (Wan et al., 2023).

4.2 Evaluation Metrics

We use BLEU (Papineni et al., 2002) and Entity F1 (Raffel et al., 2020) as evaluation metrics to assess

Models	MWOZ		SMD		CamRest	
	BLEU	Entity F1	BLEU	Entity F1	BLEU	Entity F1
DF-Net(Qin et al., 2020)	9.40	35.10	14.40	62.70	-	-
GPT-2+KE(Madotto et al., 2020)	15.05	39.58	17.35	59.78	18.00	54.85
EER(He et al., 2020b)	13.60	35.60	17.20	59.00	19.20	65.70
FG2Seq(He et al., 2020a)	14.60	36.50	16.80	61.10	20.20	66.40
CD-NET(Raghu et al., 2021)	11.90	38.70	17.80	62.90	21.80	68.60
GraphMemDialog(Wu et al., 2022)	14.90	40.20	18.80	64.50	22.30	64.40
ECO(Huang et al., 2022)	12.61	40.87	-	-	18.42	71.56
DialoKG(Rony et al., 2022)	12.60	43.50	20.00	65.90	23.40	**75.60**
UnifiedSKG(T5-Base)(Xie et al., 2022)	-	-	17.41	66.45	-	-
UnifiedSKG(T5-Large)(Xie et al., 2022)	13.69	46.04	17.27	5.86	20.31	71.03
Q-TOD(T5-base)(Tian et al., 2022)	-	-	20.14	68.22	-	-
Q-TOD(T5-Large)(Tian et al., 2022)	17. 62	50.61	21.33	71.11	23.75	74.22
ChatGPT(OpenAI, 2022)	7.47	32.87	15.29	54.71	14.60	58.11
DF-TOD(T5-Base)(Shi et al., 2023)	18.26	52.52	24.12	69.36	25.85	72.83
DF-TOD(T5-Large)(Shi et al., 2023)	18.48	53.17	25.10	71.58	26.00	74.04
MK-TOD$_{ctr}$(T5-Base)(Ding et al., 2024)	17.33	51.86	24.77	67.86	26.76	73.60
MK-TOD$_{ctr}$(T5-Large)(Ding et al., 2024)	17.55	52.97	25.43	**73.31**	26.20	71.72
Gemini(Team et al., 2023)	-	32.38	-	57.64	-	62.13
MK-TOD$_{pre}$(ChatGPT)(Ding et al., 2024)	7.22	32.78	15.07	58.41	15.56	54.96
MK-TOD$_{pro}$(ChatGPT)(Ding et al., 2024)	7.58	32.84	15.24	59.72	16.07	56.83
MAKER(T5-Base)(Wan et al., 2023)	17.23	53.68	24.79	69.79	25.04	73.09
MAKER(T5-Large)(Wan et al., 2023)	18.77	54.72	25.91	71.30	25.53	74.36
Ours(T5-Base)	19.24	54.72	26.59	70.79	27.47	73.93
Ours(T5-Large)	**19.42**	**55.72**	**27.13**	72.51	**27.49**	74.13

Table 1: Overall results of end-to-end TOD systems with dialogue-level KB on MWOZ, SMD, and CamRest. The best scores are highlighted in bold, and the second-best scores are underlined.

the quality of generated responses. BLEU measures the fluency of the generated responses. Entity F1 evaluates whether the generated responses contain the correct knowledge by computing the micro-averaged precision and recall of attribute values appearing in the outputs.

4.3 Implementation Detail

We instantiate the entity retriever using a BERT-Base (Devlin et al., 2019) model and the response generator using two variants of the T5 model (Raffel et al., 2020): T5-Base and T5-Large. All models are fine-tuned using the AdamW (Loshchilov and Hutter, 2019) optimizer with different learning rate schedulers and a batch size of 2. The retriever is trained with a fixed learning rate scheduler, while the generator adopts a linear scheduler. More detailed settings can be found in Appendix A.

4.4 Baselines

We compare CaTER with a set of baselines, including memory network-based, implicit retrieval-based, and explicit retrieval-based methods. The detailed descriptions of all baseline models are provided in Appendix B.

4.5 Main Results

This section presents the experimental results of the proposed CaTER framework in both dialogue-level and dataset-level KB settings.

Models	MWOZ		CamRest	
	BLEU	Entity F1	BLEU	Entity F1
DF-Net	6.45	27.31	-	-
EER	11.60	31.86	20.61	57.59
FG2Seq	10.74	33.68	19.20	59.35
CD-NET	10.90	31.40	16.50	63.60
Q-TOD(T5-Large)	16.67	47.13	21.44	63.88
ChatGPT	6.79	30.31	14.76	52.92
DF-TOD(T5-Base)	17.61	51.61	27.39	70.74
DF-TOD(T5-Large)	18.36	52.96	26.61	73.58
MK-TOD$_{ctr}$(T5-Base)	17.56	50.09	26.85	73.51
MK-TOD$_{ctr}$(T5-Large)	17.40	**53.26**	27.82	71.98
MK-TOD$_{pre}$(ChatGPT)	7.01	30.69	14.51	52.38
MK-TOD$_{pro}$(ChatGPT)	7.31	32.04	14.91	53.58
MAKER(T5-Base)	16.25	50.87	26.19	72.09
MAKER(T5-Large)	18.23	52.12	25.34	72.43
Ours(T5-Base)	18.35	52.43	28.89	74.29
Ours(T5-Large)	**19.11**	52.97	**29.12**	**75.65**

Table 2: Overall results of end-to-end TOD systems with dataset-level KB on MWOZ, SMD, and CamRest.

4.5.1 Dialogue-level KB

The experimental results under the dialogue-level KB setting are shown in Table 1. The proposed CaTER framework, instantiated with T5-Large, achieves state-of-the-art (SOTA) performance on both the MWOZ and SMD datasets. Specifically, compared to MAKER, CaTER improves the Entity F1 score by 1.00 on MWOZ and 1.21 on SMD. Notably, CaTER also achieves the highest BLEU scores, with gains of 0.65 on MWOZ, 1.22 on SMD, and 2.63 on CamRest over MAKER. How-

Models	T5-Base		T5-Large	
	BLEU	Entity F1	BLEU	Entity F1
$\text{Ours}_{\text{condensed}}$	19.24	54.72	19.42	55.72
w/o CyCAD	19.12	53.78	19.31	54.79
w/o TER	17.14	48.24	17.12	52.01
$\text{Ours}_{\text{full}}$	18.35	52.43	19.11	52.97
w/o CyCAD	17.02	48.20	18.31	52.37
w/o TER	16.59	48.11	18.25	51.66

Table 3: Results of Ablation study and different generator backbones, where '*w/o* CyCAD' denotes replacing the CyCAD attention mechanism in CaTER with a standard cross-attention mechanism, '*w/o* TER' denotes the removal of the TER component.

ever, CaTER does not achieve the best Entity F1 score on the CamRest dataset. This can be attributed to the limited average number of entities (1.92) per dialogue-level KB in CamRest, which poses a challenge for the retriever while favoring autoregressive models such as DialoKG and UnifiedSKG. In contrast, on the MWOZ dataset, which contains an average of more than 7 entities per KB, CaTER significantly outperforms existing SOTA methods.

Under the same generator backbone models (T5-Base and T5-Large), the proposed CaTER framework consistently outperforms methods such as MAKER and DF-TOD in the dialogue-level KB setting. The reason behind this phenomenon can be explained by CaTER's retriever, which builds upon the distillation training strategy of MAKER and further enhances entity modeling through TER. By capturing latent relationships among topology entities, the topology entity contrastive loss pulls positive samples closer to the dialogue context while pushing hard negatives farther apart. This encourages the retriever to focus on fine-grained attribute differences between entities, enabling more accurate identification of entities aligned with users' implicit intent.

Additionally, CyCAD helps the retriever concentrate more on the intrinsic attributes of entities by reducing interference from weakly relevant contextual information. Through joint distillation training between the retriever and the generator, CaTER improves the alignment between retrieved entities and generated responses.

4.5.2 Dataset-level KB

The experimental results under the dataset-level KB setting are shown in Table 2. The proposed CaTER framework demonstrates significant advantages in this scenario. When comparing the results in Table 1 and Table 2, we observe that CaTER exhibits more stable performance when handling larger KB. Specifically, under the same generator backbone, CaTER consistently achieves significant improvements over baseline methods in both BLEU and Entity F1 scores on the MWOZ and CamRest datasets. It may be attributed to CaTER's ability to capture implicit relationships among entities by modeling their relevance to the dialogue context and user intent. Such capability enhances CaTER's effectiveness in identifying the correct entities, offering generalizability across both domain-rich datasets like MWOZ and entity-diverse datasets like CamRest. These results confirm the superiority of CaTER in handling large-scale knowledge bases.

5 Analysis and Discussion

5.1 Ablation Study

We conduct ablation studies on the MWOZ dataset to evaluate the proposed CaTER framework under both dialogue-level and dataset-level KB settings. The results are presented in Table 3.

In ablation study, removing the TER module leads to substantial degradation across metrics under both knowledge-base (KB) configurations, with the drop most pronounced for Entity-F1. This suggests that TER strengthens the retriever's ability to disambiguate among closely related entities and, through attribute-score filtering, is particularly beneficial when the candidate set is small (dialogue-level KB). Likewise, replacing CyCAD diminishes performance—more markedly under the dataset-level KB—indicating that CyCAD better identifies entity information aligned with user intent while leveraging entity–context interactions to encourage more diverse responses. Collectively, these findings substantiate the effectiveness of CaTER for end-to-end task-oriented dialogue: each component contributes meaningfully, and removing any one of them consistently harms performance across both evaluation metrics.

5.2 Comparison of Retrieval Methods

We evaluate different retrieval strategies under the dataset-level KB setting on MWOZ, following (Wan et al., 2023), using T5-Base as the backbone. In addition, we report Recall@x to assess whether entities in the generated response appear among the Top-x retrieved candidates.

As shown in Figure 4, CaTER achieves the highest BLEU and Entity F1 scores, demonstrating

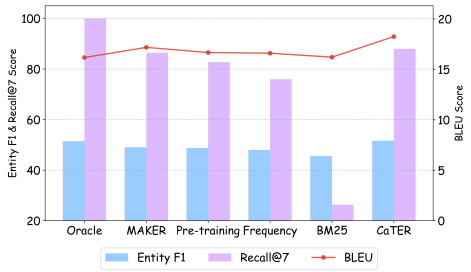

Figure 4: Comparison of different retrieval methods under the dataset-level KB setting. 'Oracle' using the dataset-level KB itself as the retrieval result for each dialogue. 'Frequency' measures relevance based on the frequency of attribute values appearing in the dialogue context. 'BM25' computes relevance based on the BM25 score between the dialogue context and each entity.

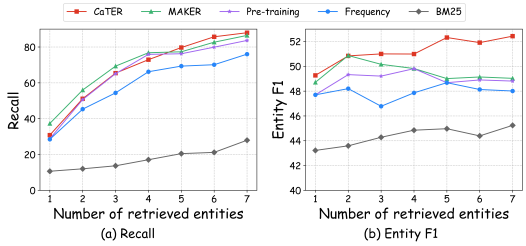

Figure 5: Recall (Figure 5(a)) and Entity F1 (Figure 5(b)) scores of different retrieval methods under the dataset-level KB setting, evaluated with varying numbers of retrieved entities.

strong performance in both response generation and entity retrieval. While it ranks second to the Oracle in Recall@7, this is expected since the Oracle retrieves the entire KB without disambiguating similar entities, resulting in lower response quality. In contrast, CaTER leverages TER to distinguish between similar entities, significantly outperforming the Oracle in BLEU and Entity F1, while maintaining competitive Recall@7 with fewer retrieved entities.

To further analyze the effect of retrieval size, we plot Recall and Entity F1 against the number of retrieved entities (Figure 5). Recall increases consistently across all methods, with CaTER surpassing MAKER beyond five candidates and achieving the highest Recall overall. However, a larger retrieval size incurs higher computational costs. Meanwhile, Entity F1 remains relatively stable or even declines with larger retrieval sizes, as additional candidates may inject noise into the generation process. Notably, CaTER surpasses MAKER's peak Entity F1 with fewer retrieved entities, indicating its efficiency in producing high-quality responses under lower overhead.

Notably, while CaTER's gain is more limited on CamRest due to its small KB size and minimal ambiguity, our goal is to improve performance in more challenging real-world settings where retrieval noise is substantial.

To assess the statistical significance of our improvements, we conducted Welch's t-test comparing our models (T5-Base and T5-Large) against strong baselines (e.g., MAKER) under both the condensed and full settings. As reported in Appendix D, the majority of the improvements in BLEU and Entity F1 across MWOZ, SMD, and CamRest are statistically significant ($p < 0.05$), validating the robustness of our method. This analysis provides further evidence that the performance gains observed are not due to random fluctuations, but stem from the effectiveness of our proposed contrastive framework and attention mechanisms.

6 Conclusion

In this paper, we propose a novel context-aware topology entity retrieval contrastive learning framework (CaTER), which incorporates a cycle context-aware distilling attention mechanism to construct semantically correlated topological hard negative samples. A topology entity contrastive loss is then used to distill knowledge back into the retriever through reverse training. Experimental results demonstrate that CaTER effectively captures latent relationships among KB entities and significantly improves the retriever's ability to select entities aligned with user intent in both dialogue-level and dataset-level KB settings. CaTER alleviates this issue by bridging the gap between retrieved entities and generated responses. Overall, this work offers a new perspective and practical solution for enhancing entity retrieval and response generation in end-to-end TOD systems, with considerable practical value for generalization and real-world application.

7 Acknowledgements

The authors look forward to the insightful comments and suggestions of the anonymous reviewers and editors, which will go a long way towards improving the quality of this paper. This work is supported by Research Projects of the Nature Science Foundation of Hebei Province (F2020402003).

Limitations

There are two potential limitations in this paper that warrant further consideration. First, the incorporation of topology entity retrieval contrastive learning increases the difficulty for the retriever to identify correct entities during training, which may reduce training efficiency. Second, this work has not yet explored alternative strategies for selecting topology hard negatives, such as approaches based on knowledge graphs or graph convolutional networks. The absence of such techniques may limit the generator's generalization ability when dealing with complex semantic structures.

References

Yucheng Cai, Hong Liu, Zhijian Ou, Yi Huang, and Junlan Feng. 2023. Knowledge-retrieval task-oriented dialog systems with semi-supervision. In *Proc. Interspeech 2023*, pages 4673–4677.

Cheng Chen, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, Yudong Zhu, and Xiaodong Gu. 2022a. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 18103–18112.

Feilong Chen, Xiuyi Chen, Shuang Xu, and Bo Xu. 2022b. Improving cross-modal understanding in visual dialog via contrastive learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7937–7941. IEEE.

Zhanpeng Chen, Zhihong Zhu, Wanshi Xu, Xianwei Zhuang, and Yuexian Zou. 2024. Relevance is a guiding light: Relevance-aware adaptive learning for end-to-end task-oriented dialogue system. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5410–5420.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Zeyuan Ding, Zhihao Yang, Ling Luo, Yuanyuan Sun, and Hongfei Lin. 2024. From retrieval to generation: A simple and unified generative model for end-to-end task-oriented dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17907–17914.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, and 1 others. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10749–10757.

Zhenhao He, Yuhong He, Qingyao Wu, and Jian Chen. 2020a. Fg2seq: Effectively encoding knowledge for end-to-end task-oriented dialog. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8029–8033. IEEE.

Zhenhao He, Jiachun Wang, and Jian Chen. 2020b. Task-oriented dialog generation with enhanced entity representation.

Guanhuan Huang, Xiaojun Quan, and Qifan Wang. 2022. Autoregressive entity generation for end-to-end task-oriented dialog. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 323–332.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-guided contrastive learning of unsupervised dialogue embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7272–7282.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394.

OpenAI. 2022. Chatgpt: Conversational ai language model. In *Website. https://openai.com/chatgpt.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with kb retriever. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 133–142.

Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Dinesh Raghu, Atishya Jain, Sachindra Joshi, and 1 others. 2021. Constraint based knowledge base distillation in end-to-end task oriented dialogs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5051–5061.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialokg: Knowledge-structure aware task-oriented dialogue generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2557–2571.

Weizhou Shen, Yingqi Gao, Canbin Huang, Fanqi Wan, Xiaojun Quan, and Wei Bi. 2023. Retrieval-generation alignment for end-to-end task-oriented dialogue system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8261–8275.

Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. 2023. Dual-feedback knowledge retrieval for task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6566–6580.

Jie Tan, Hengyi Cai, Hongshen Chen, Hong Cheng, Helen Meng, and Zhuoye Ding. 2023. Contrastive learning with dialogue attributes for neural dialogue generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xin Tian, Yingzhan Lin, Mengfei Song, Siqi Bao, Fan Wang, Huang He, Shuqi Sun, and Hua Wu. 2022. Q-tod: A query-driven task-oriented dialogue system. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7260–7271.

Fanqi Wan, Weizhou Shen, Ke Yang, Xiaojun Quan, and Wei Bi. 2023. Multi-grained knowledge retrieval for end-to-end task-oriented dialog. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11196–11210.

Jiaan Wang, Jianfeng Qu, Kexin Wang, Zhixu Li, Wen Hua, Ximing Li, and An Liu. 2024. Improving the robustness of knowledge-grounded dialogue via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19135–19143.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.

Jie Wu, Ian G Harris, and Hongzhi Zhao. 2022. Graph-memdialog: Optimizing end-to-end task-oriented dialog systems using graph memory networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11504–11512.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, and 1 others. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631.

Wanshi Xu, Xuxin Cheng, Zhihong Zhu, Zhanpeng Chen, and Yuexian Zou. 2024. Learning to match representations is better for end-to-end task-oriented dialog system. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10409–10419.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024.

Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219.

A More Implementation Details

The hyperparameters of CaTER under dialogue-level KB and dataset-level KB setting are shown in Table 4 and Table 5, respectively. We conduct all experiments on a single 24G NVIDIA RTX 4090 GPU and select the best checkpoint based on model performance on the validation set.

B Baselines Details

We compare CaTER with the following baselines, which are organized into four categories according to how they model entity retrieval.

1. **Memory Network**: This category of methods embeds the KB into a memory network, where the dialogue context representation is used as a query to retrieve relevant information and generate responses.
   (a) **DF-NET** (Qin et al., 2020): Proposed a dynamic fusion module to capture inter-domain dependencies, enabling effective handling of multi-domain dialogue tasks.
   (b) **EER** (He et al., 2020b): Proposed an enhanced entity representation approach that captures both context-sensitive and structure-aware representations of entities.
   (c) **FG2Seq** (He et al., 2020a): Encodes entity knowledge by leveraging both the inherent structural information in the knowledge graph and the latent semantic information in the dialogue context.
   (d) **CD-NET** (Raghu et al., 2021): Disentangles context-independent KB information by incorporating paired similarity filters and auxiliary loss functions, enabling effective TOD response generation.
   (e) **GraphMemDialog** (Wu et al., 2022): Introduced a novel graph-based memory-augmented Seq2Seq architecture that learns structural patterns in the dialogue context and generators dynamic interactions between the dialogue and the KB.
2. **Implicit Retrieval**: This category of methods embeds the KB into generator parameters by data augmentation, allowing the generator to perform implicit retrieval during response generation.
   (a) **GPT-2+KE** (Madotto et al., 2020): Transforms the KB into equivalent knowledge embeddings and integrates them into the dialogue, effectively storing the KB within the generator parameters.
   (b) **ECO** (Huang et al., 2022): Enforces entity consistency in generated responses by using trie-constrained autoregressive decoding to select the most relevant entities.
3. **Explicit Retrieval**: This category of methods separates entity retrieval from response generation, explicitly retrieving relevant entity information from the KB based on user intent to guide response generation.
   (a) **DialoKG** (Rony et al., 2022): Proposed a TOD system that leverages the structural information of a knowledge graph to enhance reasoning ability, effectively integrating knowledge into the generator.
   (b) **UnifiedSKG** (Xie et al., 2022): Proposed a unified framework for structured knowledge grounding that utilizes both KB-based semantic parsing and knowledge-based question answering data to fulfill user requests.
   (c) **Q-TOD** (Tian et al., 2022): Separates knowledge retrieval from response generation by extracting key information from the dialogue context and retrieving relevant knowledge records to generate appropriate responses.
   (d) **DF-TOD** (Shi et al., 2023): Trains the retriever using pseudo-labels obtained from the generator feedback, improving retrieval quality without requiring additional supervision.
   (e) **MK-TOD** (Ding et al., 2024): Incorporates retrieved entities and meta-knowledge to guide generator training,

Hyperparameters	MWOZ		SMD		CamRest	
	T5-Base	**T5-Large**	**T5-Base**	**T5-Large**	**T5-Base**	**T5-Large**
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Batch Size	2	1	2	1	2	1
Gradient Accumulation Steps	32	64	32	64	32	64
Learning Rate Schedule	Linear	Linear	Linear	Linear	Linear	Linear
Learning Rate	2e-5	1e-4	1e-4	1e-4	1e-4	7e-5
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01
Entity Encoder Max Length	128	128	128	128	128	128
Generator Max Context Length	200	200	200	200	200	200
Generator Max KB Length	100	100	100	100	100	100
Max Response Length	64	64	128	128	64	64
Top-K Retrieval Entities	6	7	8	8	6	4
Attribute Score threshold	0.10	0.05	0.00	0.00	0.00	0.00
Distillation Start Gradient Steps	20000	60000	27000	30000	24000	29000

Table 4: The hyperparameter settings of CaTER under dialogue-level KB are used on the MWOZ, SMD, and CamRest datasets.

Hyperparameters	MWOZ		CamRest	
	T5-Base	**T5-Large**	**T5-Base**	**T5-Large**
Optimizer	AdamW	AdamW	AdamW	AdamW
Batch Size	2	1	2	1
Gradient Accumulation Steps	32	64	32	64
Learning Rate Schedule	Linear	Linear	Linear	Linear
Learning Rate	2e-5	1e-4	1e-4	1e-4
Weight Decay	0.01	0.01	0.01	0.01
Entity Encoder Max Length	128	128	128	128
Generator Max Context Length	200	200	200	200
Generator Max KB Length	100	100	100	100
Max Response Length	64	64	64	64
Top-K Retrieval Entities	10	10	10	10
Attribute Score threshold	0.20	0.20	0.20	0.20
Distillation Start Gradient Steps	36000	120000	30000	60000

Table 5: The hyperparameter settings of CaTER under dataset-level KB are used on the MWOZ and CamRest datasets.

User Utterance	i am looking for a restaurant. the restaurant should serve indian food and should be in the expensive price range.						
Retrieval Entities	**Name**	**Address**	**Area**	**Food**	**Phone**	**Postcode**	**Price**
	curry garden	~~106 regent street city centre~~	~~centre~~	indian	~~01223302330~~	~~cb21dp~~	expensive
	sitar tandoori	~~43 high street cherry hinton~~	~~east~~	indian	~~01223249955~~	~~cb19hx~~	expensive
CaTER Response	curry garden is an expensive indian restaurant. is there anything else i can help you with?						
Golden Response	curry garden is a great restaurant that fulfills your wishes. it is located at 106 regent street city centre.						
User Utterance	awesome. can i also get the area, post code, and phone number?						
Retrieval Entities	**Name**	**Address**	**Area**	**Food**	**Phone**	**Postcode**	**Price**
	curry garden	~~106 regent street city centre~~	centre	~~indian~~	01223302330	cb21dp	~~expensive~~
	sitar tandoori	~~43 high street cherry hinton~~	east	~~indian~~	01223249955	cb19hx	~~expensive~~
CaTER Response	sure, the curry garden is located in the centre part of town, the postcode is cb21dp, and the phone number is 01223302330.						
Golden Response	yes, curry garden is located in the centre and the postcode is cb21dp.						

Table 6: A dialogue case from the MWOZ dataset demonstrating the effectiveness of CaTER. Blue text highlights knowledge-grounded entities.

User Utterance	hi, i'm looking for a moderately priced restaurant that serves chinese food.						
Retrieval Entities	**Name**	**Address**	**Area**	**Food**	**Phone**	**Postcode**	**Price**
	golden palace	~~25 king's parade~~	~~centre~~	chinese	~~01223334567~~	~~cb21hs~~	moderate
	wok express	~~12 milton road~~	~~north~~	chinese	~~01223331234~~	~~cb18yy~~	moderate
CaTER Response	sure, golden palace is a moderately priced chinese restaurant. is there anything else you need?						
Golden Response	golden palace is a moderately priced chinese restaurant located in the centre. its address is 25 king's parade, and the postcode is cb21hs.						

Table 7: A failure dialogue case from the CamRest dataset demonstrating the effectiveness of CaTER. Blue text highlights knowledge-grounded entities.

improving the utilization of external KB in dialogue response generation.

(f) **MAKER** (Wan et al., 2023): Constructs a multi-grained knowledge retriever based on entity and attribute score to better align retrieved entity with user intent.

4. **Few-shot LLM-based**: This category of methods generate responses by prompting large language models with limited exemplars and retrieval context, without parameter updates.

(a) **ChatGPT** (OpenAI, 2022): As a large language models, ChatGPT demonstrates strong performance in human dialogue. In this paper, we construct a baseline by using ChatGPT (GPT-3.5-turbo API) as the response generator.

(b) **Gemini** (Team et al., 2023): Developed by Google, is a state-of-the-art large language model capable of generating contextually coherent responses.

(c) **MK-TOD**$_{pre}$ (Ding et al., 2024): Employs a simple few-shot prompting strategy over ChatGPT (GPT-3.5-turbo), where the model generates responses based on a small number of exemplars and retrieved knowledge, without specialized reasoning prompts or task decomposition.

(d) **MK-TOD**$_{pro}$ (Ding et al., 2024): Enhances the above setting by incorporating structured prompting templates that guide ChatGPT to perform explicit belief state tracking and entity selection before response generation, improving task-awareness and grounding consistency.

C Case Study

A successful dialogue example from the MWOZ dataset are showed in Table 6. It can be observed that, for a given user query, the proposed CaTER framework successfully retrieves entities that align with the user's intent, while masking weakly relevant attribute features. Furthermore, when the user intent shifts, as in the second turn of the example, where the user requests information such as address, postcode, and phone number, CaTER dynamically re-evaluates the relevance scores of the attributes and generates a response containing the corresponding information.

We present a failure case from the restaurant domain in the SMD dataset, where CaTER produces a fluent but partially incorrect response due to entity-level confusion in the retrieval stage, as showed in Table 7. Despite correctly identifying golden palace in the first turn, CaTER hallucinates all follow-up attributes, such as postcode, area, and address, from a hard negative (wok express). This

indicates a severe attribute contamination problem, where the model fails to isolate grounded fields even after confident entity selection. The error likely stems from: 1. Shared attributes (food = chinese, price = moderate) between top entities. 2. Incomplete attribute alignment in contrastive training. 3. CyCAD's attention not sufficiently disambiguating slot-level details in multi-attribute responses.

These cases demonstrates that CaTER may correctly retrieve the target entity but still suffer from attribute leakage in downstream generation. It highlights the need for slot-level grounding supervision, which we plan to explore by incorporating attribute-specific disentanglement or contrastive penalties during training in the future research.

D Statistical Significance Test Result

To evaluate whether the performance improvements of our method are statistically significant, we simulate 3 independent runs assuming a normal distribution $N(\mu, \sigma^2)$ with $\sigma = 1.0$, for both BLEU and Entity F1 scores. In order to effectively evaluate the validity of the CaTER model in condensed scenarios, explicit retrieval methods and the LLM few-shot methods are selected for t-test evaluation. We then perform Welch's t-test between our models (CaTER(T5-Base) and CaTER(T5-Large)) and the strongest baselines. Tables 8 and 9 report the mean, standard deviation (SD), t-statistics, and p-values under both Condensed and Full settings. Most comparisons yield p-values below 0.05, confirming the significance of our improvements. We clarify that: 1. In the Full scenario, significance is measured against all the baseline models.

2. In the Condensed scenario, we restrict comparison to strong implicit retrieval baselines and LLM-based few-shot methods, as these represent the most competitive paradigms under limited supervision and reduced KB complexity.

This distinction ensures a fair and context-aware comparison tailored to the scope and assumptions of each evaluation protocol.

E Notation and Definitions

Fig **??** summarizes the main notations and selected parameters used in the methodology section (Section 3). The table is intended to provide readers with a concise reference to the symbols, variables, and configuration details that appear throughout the proposed approach. Only the most relevant items are included, focusing on those essential for understanding the core components of the method.

Models	MWOZ		SMD		CamRest	
	BLEU	**Entity F1**	**BLEU**	**Entity F1**	**BLEU**	**Entity F1**
Ours(T5-Base)	19.24	54.72	26.59	70.79	27.47	73.93
Ours(T5-Large)	19.42	55.72	27.13	72.51	27.49	74.13
mean	14.75	45.49	21.70	65.89	22.54	66.89
SD	4.82	9.71	4.42	6.48	4.74	7.22
t(T5-Base)	2.94	3.15	3.50	2.51	3.29	2.67
p(T5-Base)	0.0164	0.0103	0.0067	0.0311	0.0093	0.0235
t(T5-Large)	3.06	3.50	3.89	3.39	3.31	2.76
p(T5-Large)	0.0135	0.0058	0.0037	0.0069	0.0091	0.0203

Table 8: Performance Comparison with Baseline Models, in the Condensed scenario: Mean, Standard Deviation(SD), and Statistical Significance (t/p-values)

Models	MWOZ		CamRest	
	BLEU	**Entity F1**	**BLEU**	**Entity F1**
Ours(T5-Base)	18.35	52.43	28.89	74.29
Ours(T5-Large)	19.11	52.97	29.12	75.65
mean	13.06	41.10	21.69	64.43
SD	4.65	10.23	5.07	8.11
t(T5-Base)	4.97	4.98	3.47	2.83
p(T5-Base)	0.0003	0.0002	0.0046	0.0153
t(T5-Large)	5.12	5.35	3.86	3.59
p(T5-Large)	0.0002	0.0001	0.0023	0.0037

Table 9: Performance Comparison with Baseline Models, in the full scenario: Mean, Standard Deviation(SD), and Statistical Significance (t/p-values)

Symbol	Mean
$\mathcal{D}$	dialogue history
T	dialogue turn
$\mathcal{R}$	response
$\mathcal{K}$	external entity knowledge base
$\hat{\mathcal{K}}$	candidate entity set
$\tilde{\mathcal{K}}$	attribute candidate entity set
σ	the Sigmoid function
α	context-aware pooling weight parameter
r	the length of $\mathcal{R}$
ϵ	the bias of $Pool^c_{t,i}$
u_i	i-th user utterance
y_i	i-th system response
e_i	i-th entity in the $\mathcal{K}$
$\tilde{e}_k$	k-th entity in the $\tilde{\mathcal{K}}$
c_t	dialogue context at the t-th turn
L_c	the length of c_t
Att_c	the cross attention score
τ_a	attribute threshold
τ_t	temperature parameter
$\{a^n, v^n_i\}$	n-th attribute-value of e_i
Enc_c	context encoder
Enc_e	entity encoder
Enc_g	the encoder of generator
Dec_g	the decoder of generator
$s_{t,i}$	the entity score of e_i
$\mathbf{s}_t$	the similarity score
$\mathbf{a}_{t,i}$	the attribute score of e_i
$\mathbf{b}_t$	whether masked attribute value both in the $\tilde{\mathcal{K}}$ and c_t
$\mathbf{H}_{t,i}$	global representation of current dialogue turn
$\mathbf{H}_t$	the input of Dec_g
$Pool^c_{t,i}$	context-aware average pooling
$Pool^e_{t,i}$	context-independent entity pooling
$\mathbf{M}_{mask}$	the input attention mask matrix
s_{att}	CyCAD score
$\hat{\mathbf{S}}_{att}$	CyCAD distribution of entities
$S_{att}(\mathcal{R}, e_i)$	the entity relevance score
$P(\mathcal{R}_{t,r})$	response token probability distribution
$\mathcal{R}_{t,i}$	the i-th token in the generated response at t-turn
$A_{e_i}(\mathcal{R}, e_i)$	joint entity-attribute score
d^+	positive samples
$\mathcal{N}_{hard}$	hard negative samples
$Top-Q_{d \neq d^+}$	Top-Q samples d differ from d^+
$\mathcal{L}_{attr}$	context-aware attribute score loss
$\mathcal{L}_{gen}$	the loss of generator
$\mathcal{L}_{cl}$	topology entity contrastive loss
$\mathcal{L}_{ent}$	the selection loss
$\mathcal{L}_{total}$	the overall loss of the dialogue system

Table 10: The overall notations and parameters.